날씨 빅데이터 콘테스트

2-Stage 모델을 활용한 안개 발생 예측

TEAM 갈라진 하늘 | 조성우 김다민 이정환 정승민

2024. 08. 07.

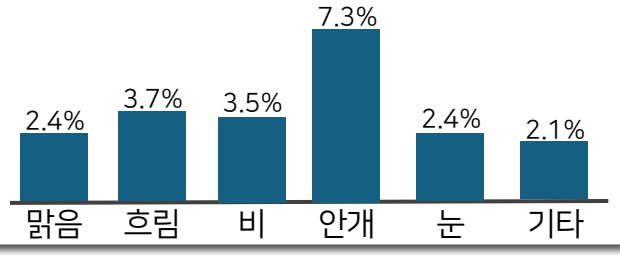
Contents

안개의 정의

대기 중의 수증기가 모여 발생하는 기상현상으로
매우 미세한 물방울이 대기 중에 떠 있어
수평시정이 1km 미만으로 감소하는 현상

교통사고 치사율 증가

<3년간('10~'12년) 초겨울(11~12월) 기상상태별 교통사고 치사율>



(권오황, 2013.11.13.)

비행기 회항을 야기

인천공항에 밤부터 안개가 짙게 끼는 바람에 도착 및 출발 항공편이 대거 결항되는 등 운행에 차질이 빚어졌다

…

지연된 항공편은 오후 10시 기준으로 출발·도착 합해 **137편**에 달했다

(연합뉴스, 2016.02.13.)

안개로 인해 많은 **인명피해**뿐만 아니라 막대한 **경제적 손실**도 발생하므로
정확한 안개 예측 모델이 필요함

안개 예측의 어려움

- 안개는 원인에 따라 **종류가 다양**하며 **시공간의 영향**을 받음
- 위성분석은 안개와 낮은 층운형 구름을 구별하기 어렵고 야간 탐지가 힘듦
- 안개가 발생하지 않은 시점이 발생 시점보다 훨씬 많은 **데이터 불균형** 상태

출처: 주요 안개분석 기술과 예측 방법, 기상청, 2015.03.18

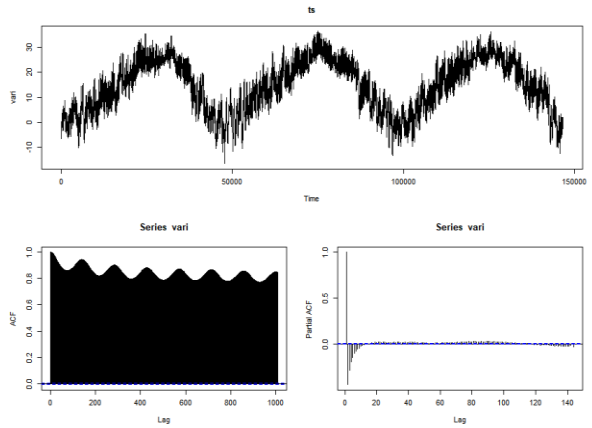
GOAL

도메인 지식을 활용하여 다양한 **파생변수들을 생성**하고
2-Stage 모델을 지역별로 적용하여 안개 발생 패턴을 예측

데이터 형태 및 특성

	year	month	day	time	minute	stn_id	ws10_deg	ws10_ms	ta	re	hm	sun10	ts	vis1	class
	I	1	1	0	10	AA	0	0	-6.4	0	38.9	0	-2.8	20000	4
	I	1	1	0	10	AA	0	0	-6.3	0	37.9	0	-2.7	20000	4
	I	1	1	0	10	AA	0	0	-6.3	0	40	0	-2.6	20000	4
								...							
	K	12	31	23	40	EC	270.3	4.6	2.1	0	51.7	0	-1.6	20000	4
	K	12	31	23	50	EC	254.8	4.1	2.1	0	53.3	0	-1.7	20000	4

시계열 특성



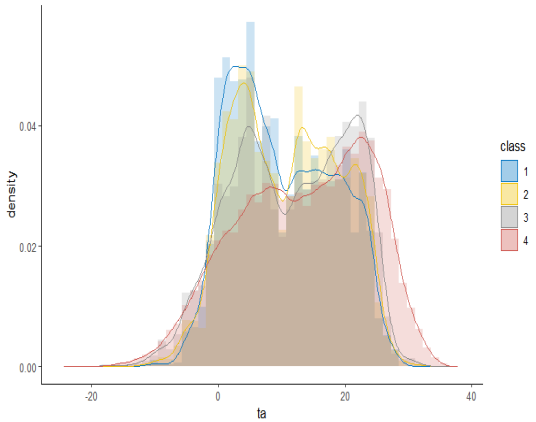
AA 지점의 ts시계열 plot, ACF,PACF

- 횡단면과 시계열 요소가 결합된 **패널 데이터**
- 매우 먼 시점까지 **자기상관성**이 존재
- 뚜렷한 **계절성**을 띔

기타 특성

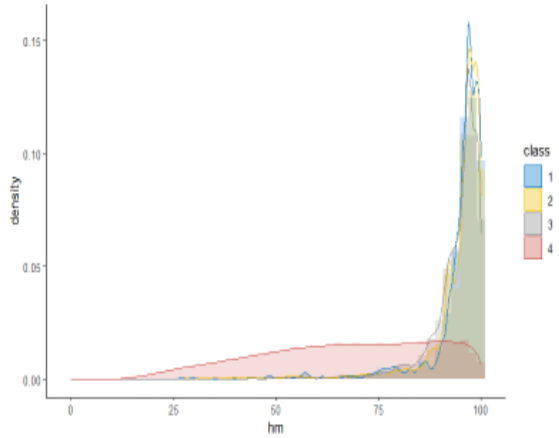
- 본 데이터는 시정거리와 시정등급 중 선택하여 회귀 혹은 분류 문제로 접근할 수 있음
  - ▶ 회귀로 접근 시 **시정거리의 분산 문제**로 모델 학습이 어려움
- 결측이 다수 발생하여 **시계열 특성을 반영한 보간**이 필요

주요 변수 Class별 EDA



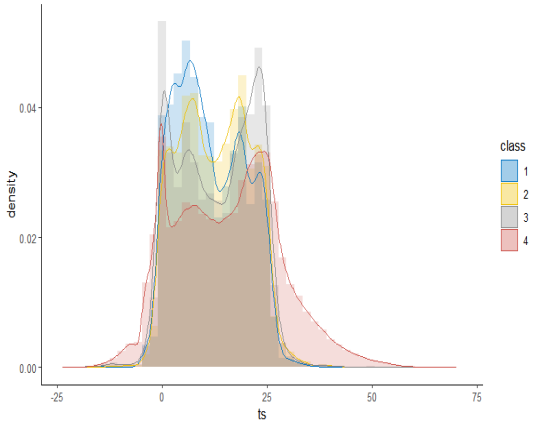
ta

상대적으로 낮은 평균 온도에서
안개 발생 시점이 밀집되어 있음



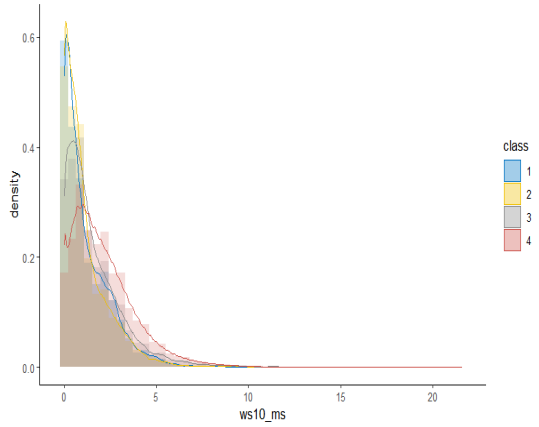
hm

습도가 매우 높은 구간에
안개 발생 시점이 밀집되어 있음



ts

지면 온도 0°C와 25 °C 구간에서
안개 발생 시점이 밀집되어 있음

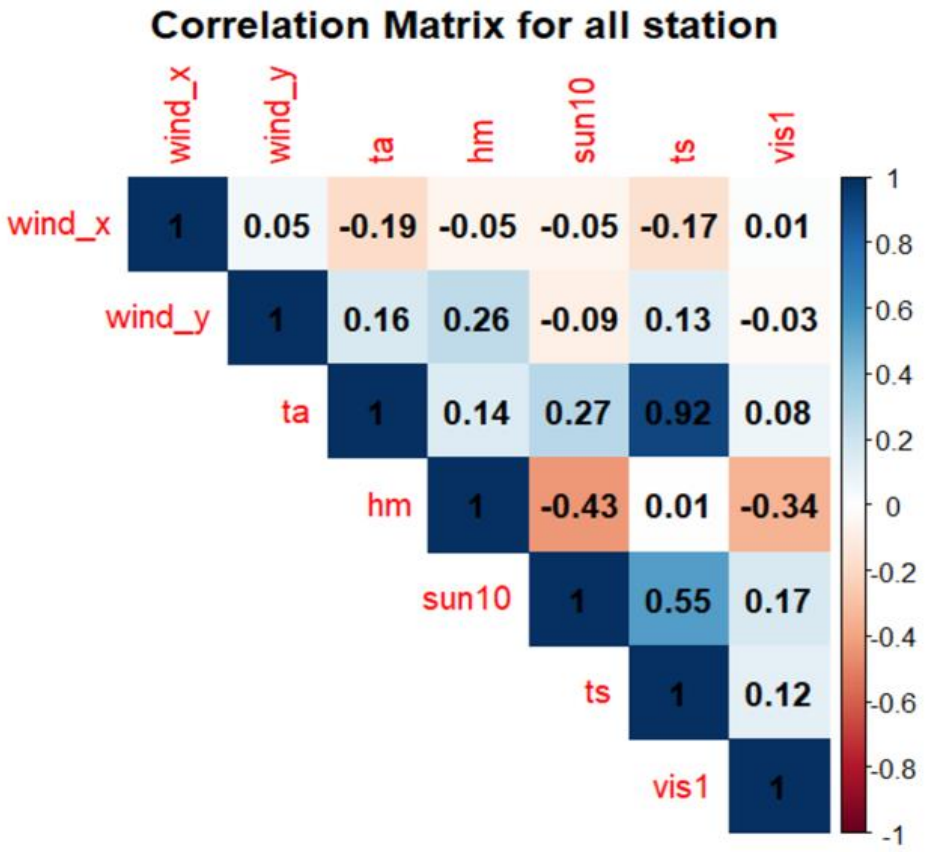


ws10_ms

풍속이 낮을 때
시정거리가 짧은 경향이 있음

전반적으로 안개 유무의 분포 차이는 뚜렷하지만, **안개의 정도는 구별하기 어려움**

변수 간 상관관계 분석



주요 상관관계

- 평균 기온(ta)과 지면 온도(ts)간 높은 상관관계 존재 (0.92)
- 상대습도(hm)와 일사량(sun10)간 유의한 음의 상관계수 확인 (-0.43)
- 시정거리(vis1)와 가장 큰 상관계수를 가진 변수는 상대습도(hm)로 습도가 높을수록 시정거리가 감소하는 경향을 확인

상관관계를 통해 얻은 인사이트를 **결측치 보간**에 사용

sun10, ts 결측치 보간

sun10

CA 지역의 장기간 연속 결측치를
인접지역인 CB 지역의 동일기간 일사량으로 대체

ts

나머지 변수의 조합으로 적합한
선형 회귀 모델($R^2 = 0.95$)을 통해 결측치 보간

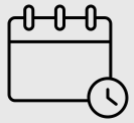
선형 보간법

연속형 변수

시계열 데이터는 **연속적**이므로, 결측 구간의 이전 시점과 이후 시점을 선형으로 연결하여 결측치를 보간
단, 변동계수를 확인하여 선형 보간의 구간 길이를 변수별로 다르게 설정
(ws10_ms, ws10_deg, sun10은 6, hm은 25, ta는 19로 설정)

시점 파생변수

cosMD,sinMD



month와 day를 결합하고, 삼각변환하여 생성

cosTM, sinTM

time과 minute을 결합하고, 삼각변환하여 생성

이슬점 관련 파생변수

dew

습도와 평균 온도를 활용하여 만든 이슬점의 온도

sup_dew

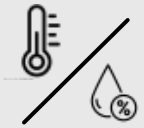
기온과 이슬점의 대소를 나타내는 이진 변수

기타 파생변수

ta-ts

평균 온도와 지면 온도의 차이

ta_lnhm



기온과 로그 상대습도 간의 비율

hmws10

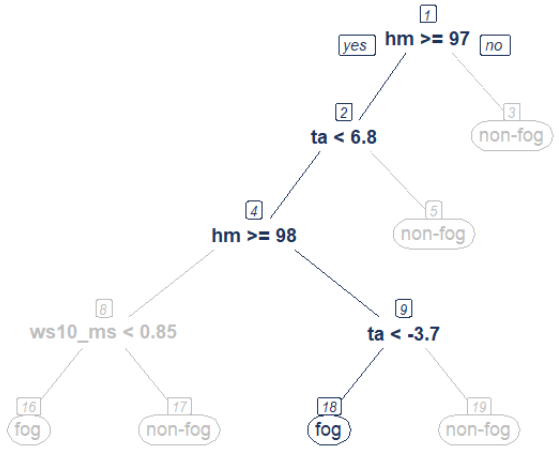
습도와 풍속의 교호작용을 반영

half

12개월을 4~10월, 11~3월로 분리한 이진 변수

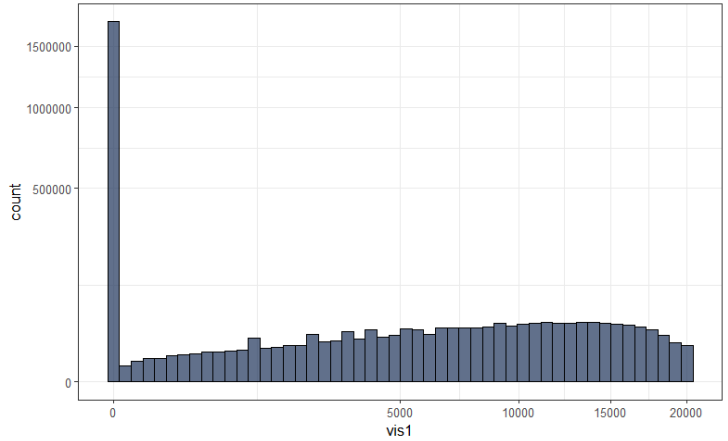
이론적 배경

안개 발생 DGP (Data Generating Process)



- 의사결정나무 기반으로 DGP를 확인해 보았을 때,
  안개 발생은 특정 기온, 습도, 풍속 등이 전부 성립해야 발생함
- 따라서 안개 발생과 미발생은 다른 DGP를 가짐. 모델링 과정에 이를 반영할 필요가 있음

Hurdle Model

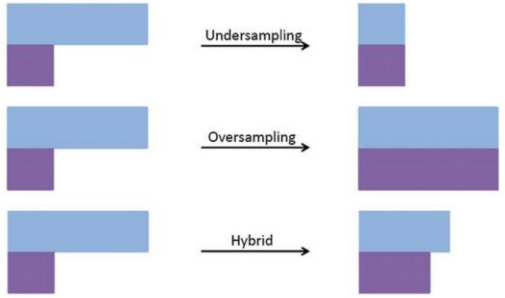


- 시정거리의 최대값을 0으로 설정하면, 상당수가 0에 속함
- 이런 영과잉 문제를 해결하기 위해 Hurdle 모델을 활용함
- 따라서 안개 발생 정도가 0일 사건(1st Stage)을 모델링하고
  안개 발생 정도(2nd Stage)에 따른 구간 분류 모델링을 진행

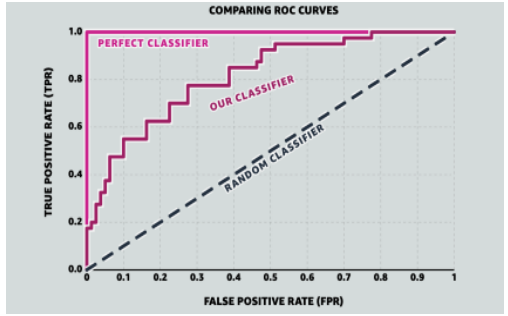
데이터 불균형

Data Imbalanced

- 안개 발생 여부에 따른 데이터 불균형 문제가 발생
- 데이터 불균형은 크게 두가지 해법을 고려할 수 있음



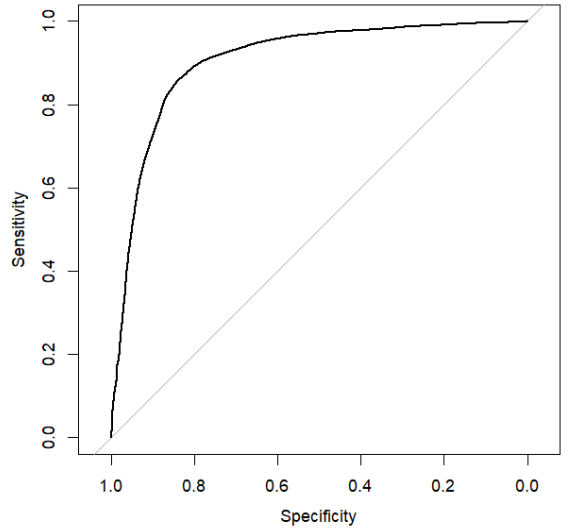
샘플링 기법



알고리즘 기법

- Smote와 같은 샘플링 기법은 시계열 특성을 해치므로 알고리즘 기법 중 Alternate Cut-off를 활용함
- 안개 발생의 판단 기준 확률을 0.5가 아닌 0.2, 0.3 등으로 바꾸는 것

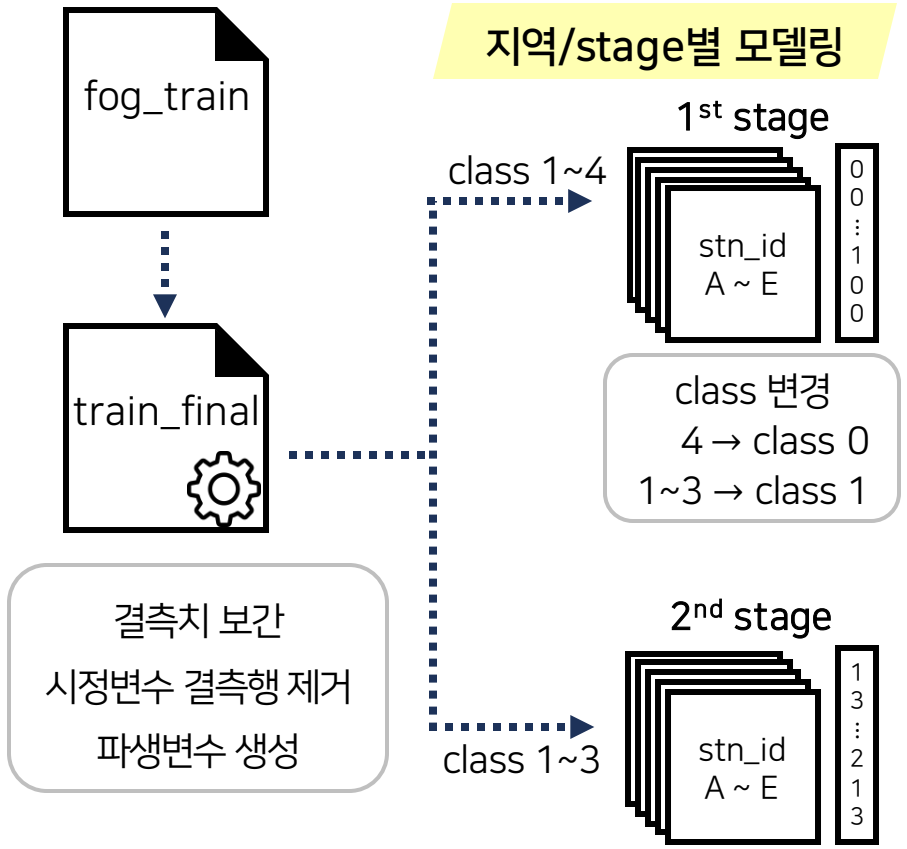
Alternate Cut-off



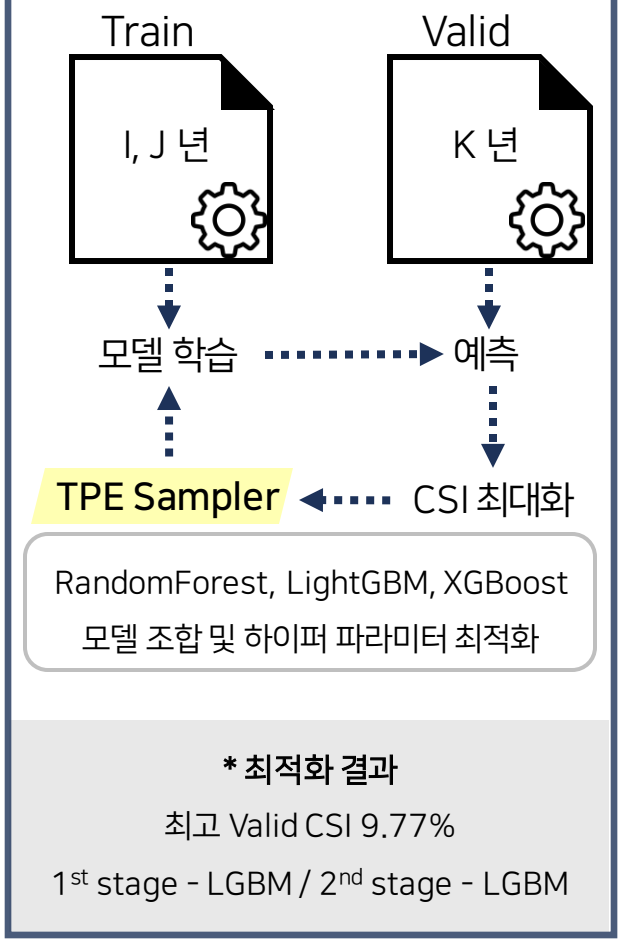
- F1-score가 최대가 되는 지점으로 Cut-off를 설정하고자 함
- Optuna를 통해 모델이 최적의 Cut-off 지점을 스스로 찾도록 함

모델링 과정

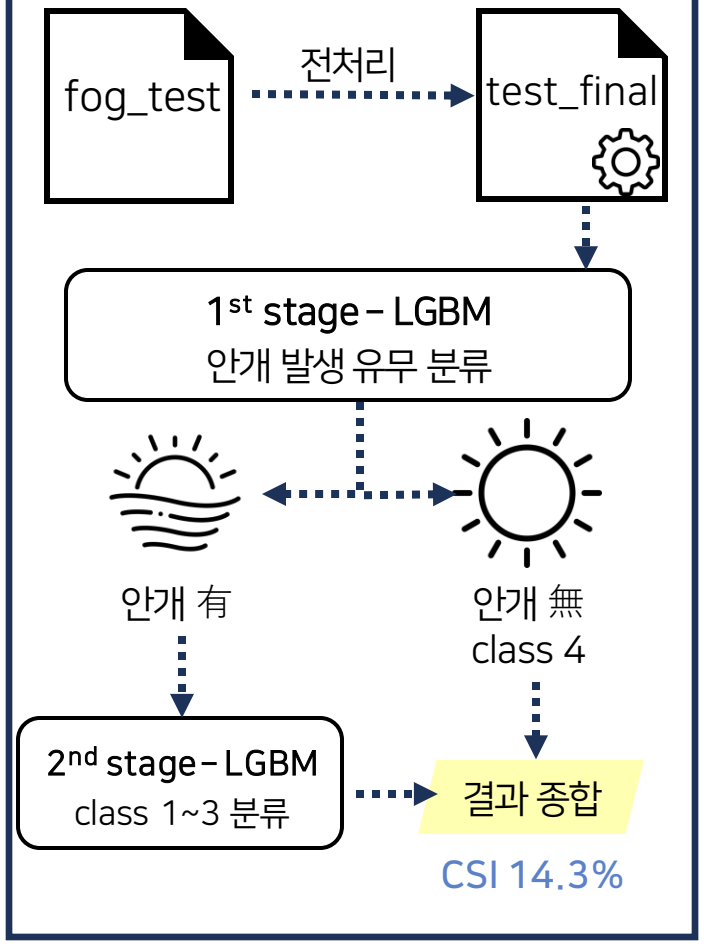
데이터 전처리



하이퍼 파라미터 최적화



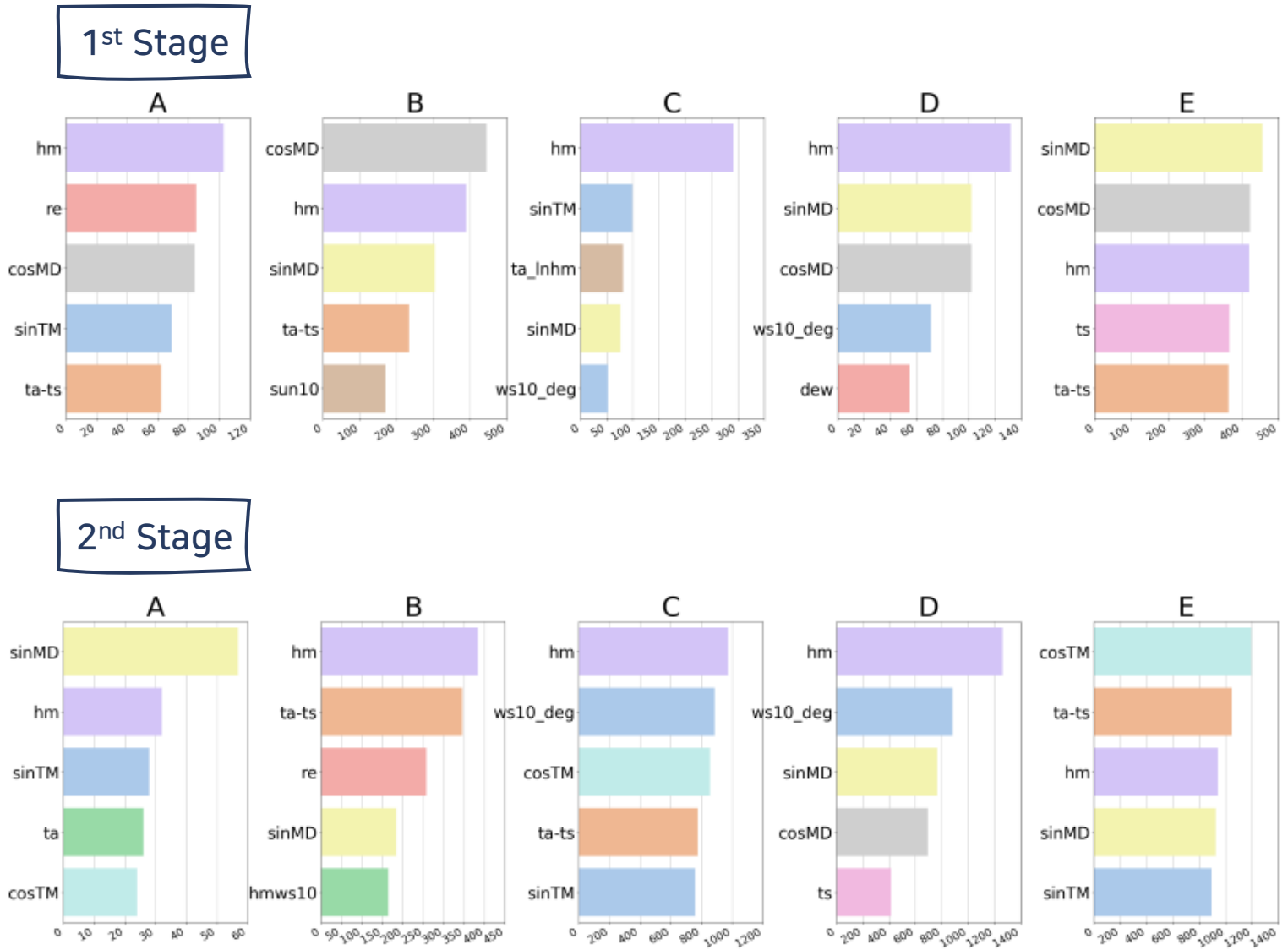
2-stage 예측 과정



변수중요도 해석 (상위 5개)

- 습도는 모든 지역과 단계에서 매우 유의했음
- ta-ts, ta_lnhm, hmws10 등 생성한 파생변수가 유의했음
- 삼각변환을 통해 생성한 시점 변수들이 매우 유의했음
- 지역별로 변수중요도 차이가 있었음

  -> 모델 학습에 지역 구분이 유효했음



분석 요약 및 의의

- 기존 연구를 통해 안개 발생 원인과 안개 예측 필요성 및 어려움 파악
- 변동계수를 이용한 선형보간법, 선형회귀모형등 변수마다 결측치 보간 기준을 설정
- 도메인 및 EDA를 기반으로 파생 변수 생성
- 클래스 불균형 해소를 위해 Hurdle 모델을 차용
- 2-Stage 모델 개발 및 예측 성능 향상
- 지리적 특성을 고려하기 위해 지역별 모델 학습

발전 가능성

- 예측시점 직전의 정보를 사용할 수 있다면 강력한 파생변수 생성 및 안정적인 결측치 보간 가능
- 지역별 안개 발생 패턴이 다르기 때문에 구체적인 지리 정보가 제공된다면 모델의 예측 성능 향상 기대